Stand Up and Move: Benchmarking Interactive Spatial Intelligence in WalkerBench

Zhiqi Ge[1*], Gang Yang[2*], Ziyang Pan[3], Jingzhe Zhu[1], Yuancheng Gu[4], Juncheng Li[1†], Qizhou Wang[5], Rui Tang[6], Siliang Tang[1†], Jun Xiao[1], and Yueting Zhuang[1†]

[1] Zhejiang University, Hangzhou, China
[2] University of Nottingham, Nottingham, UK
[3] Sun Yat-sen University, Guangzhou, China
[4] Imperial College London, London, UK
[5] Unitree Robotics, Hangzhou, China
[6] Manycore Tech Inc., Hangzhou, China

[*] Equal contribution.
[†] Corresponding authors: {junchengli, siliang, yzhuang}@zju.edu.cn

Abstract. The ultimate goal of spatial intelligence is to enable embodied agents to actively interact with the physical world. However, existing benchmarks adopt a "Spectator View"—relying on single images or passive video—neglecting the necessity of resolving geometric ambiguity through active viewpoint transformation. To address this, we present WalkerBench, a global-scale interactive benchmark built on real-world street-view and map data spanning 161 cities, with two categories of hierarchically structured spatial tasks. Our empirical evaluation reveals a critical "representation misalignment": VLMs' one-dimensional linear context structures are fundamentally incompatible with the topological nature of 3D environments, causing inevitable spatial forgetting during navigation. To overcome this, we propose Spatial-IDE, which breaks from linear dialogue history via two mechanisms: (1) *State Externalization*, transforming implicit observations into an Explicit Topological Memory (ETM); and (2) *Cognitive Decoupling*, disentangling visual understanding into Goal-Directed Perception and Spatial Reasoning, letting the VLM focus on pure high-level decision-making. Spatial-IDE achieves substantial improvements on WalkerBench and successfully generalizes zero-shot to real-world urban navigation on the Unitree humanoid robot. The project is available on `https://github.com/lalayang123456-ctrl/WalkerBench`.

1 Introduction

The ultimate frontier of AI lies in *Embodied Interaction* [12, 13, 26] within the physical world—demanding a shift from semantic pattern recognition over frozen images to purposeful engagement with dynamic 3D environments. As illustrated in Figure 1, a static first-person view is geometrically degenerate: a VLM estimating distance from a single frame hallucinates scale from unreliable priors

(300m vs. the true 500m), yet a deliberate lateral displacement immediately resolves the ambiguity via parallax. This is the essence of spatial cognition [31].

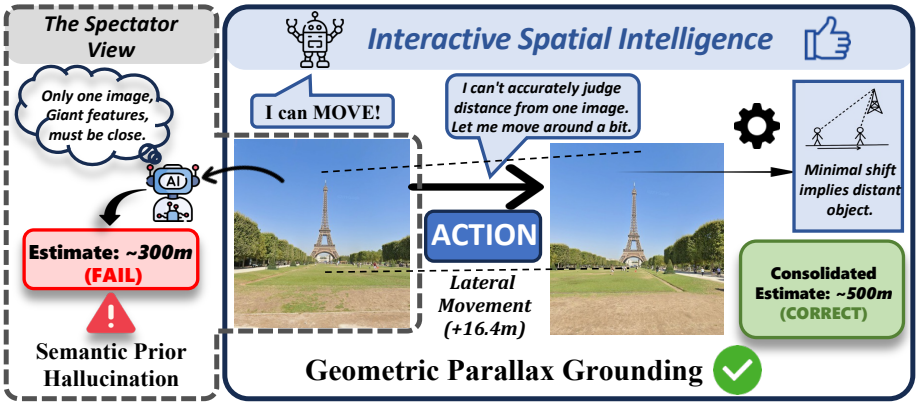

Fig. 1: Static perception versus interactive spatial intelligence. A VLM estimating distance from a single frame hallucinates scale via semantic priors (∼300m; incorrect), whereas a deliberate lateral displacement of 16.4m resolves the ambiguity through geometric parallax, yielding an accurate estimate (∼500m).

Despite growing consensus on embodied intelligence [10, 21], existing spatial benchmarks overwhelmingly adopt a *Spectator View*—evaluating from single images [6, 15, 17, 23, 30, 34] or passive video [8, 16, 31, 37], reducing agents to passive observers. Without the freedom to move, agents cannot resolve geometric ambiguity through parallax or build consistent maps—measuring only static visual commonsense rather than genuine *Interactive Spatial Intelligence*.

To address this, we introduce **WalkerBench**, a large-scale interactive benchmark built from real-world street-view imagery spanning **161 cities across 6 continents**. Unlike prior environments confined to synthetic indoor scenes [2, 14, 22] or limited outdoor settings [4, 18, 19], WalkerBench requires agents to navigate real urban topologies with only egocentric RGB observations—no GPS, no depth, no bird's-eye view. Two task families are defined: *Active Perception* (resolving geometric ambiguities through movement) and *Spatial Navigation* (long-horizon planning across eight urban topology archetypes).

Evaluation of state-of-the-art VLMs [1, 3, 9, 11, 25, 27] reveals a stark collapse: humans achieve **69.3%** while the best models reach only **24.5%**, degrading sharply with exploration depth. We term this failure *"Anterograde Spatial Amnesia"*: strong local perception at each timestep yet catastrophic loss of global coherence over time. The root cause is a *Representation Mismatch*—VLMs process context as a flat 1D token stream, wholly incommensurate with 3D non-Euclidean environments—yielding three compounding failure modes: mixed perception/memory/planning concerns, landmark dilution by growing history, and irrecoverable spatial forgetting.

Table 1: Comparison with Existing Benchmarks. WalkerBench unifies perception and navigation tasks in a global-scale interactive environment.

Category	Paradigm	Scale	Annotation	Task
Existing SI Benchmark [6, 8, 15–17, 23, 30, 37]	Static	∼k Videos	Mixed	Perception
Indoor VLN Benchmark [2, 7, 13, 14]	Interactive	∼1k Houses	Manual	Navigation
StreetLearn Series [4, 18, 19]	Interactive	2 Cities	Manual	Navigation
WalkerBench	**Interactive**	**161 Cities**	**Auto**	**Both**

To resolve this, we propose **Spatial-IDE**, a plug-and-play *Cognitive Scaffold* [24] requiring no retraining. It operates via two mechanisms: *State Externalization* maintains an **Explicit Topological Memory (ETM)**—a structured graph encoding node coordinates, orientations, and traversal history, serialized into a compact fixed-token state (`[POSITION]`, `[PATH]`, `[OBSERVATION LOG]`) that replaces raw dialogue history entirely; and *Cognitive Decoupling* [33] separates the VLM into a *Goal-Directed Perception* module for query-driven scene analysis and a *Spatial Reasoning Engine* for high-level planning over the ETM alone. Finally, we deploy Spatial-IDE **zero-shot** on a Unitree G1 humanoid robot in real urban environments, achieving kilometer-scale navigation without environment-specific fine-tuning—confirming that explicit topological memory transfers directly from benchmark to physical deployment.
Our contributions are threefold:

- **WalkerBench**: the first globally representative interactive benchmark for embodied spatial intelligence, spanning 161 cities with two task families and eight topology archetypes.
- **Diagnosis**: We formally characterize *Representation Mismatch*—the incommensurability between 1D autoregressive context and 3D topological task requirements—supported by systematic ablations across model families and topology types.
- **Spatial-IDE**: a training-free cognitive scaffold combining ETM-based State Externalization and Cognitive Decoupling, validated by significant gains on WalkerBench and zero-shot Sim-to-Real transfer on a humanoid robot.

2 Related Work

We situate **WalkerBench** within the broader landscape of spatial intelligence and embodied AI along three key dimensions (Table 1).

Spatial Intelligence Benchmarks. Static benchmarks such as VSR [17], Spatial-Bench [30], and SpatialRGPT [6] test perception within single images, while video-based counterparts [8, 15, 16, 37] add temporal dynamics from pre-recorded streams. All paradigms treat the agent as a *fixed spectator* [31]. Ours shifts evaluation to *interactive reasoning*, where the agent must actively move to resolve ambiguity via parallax.

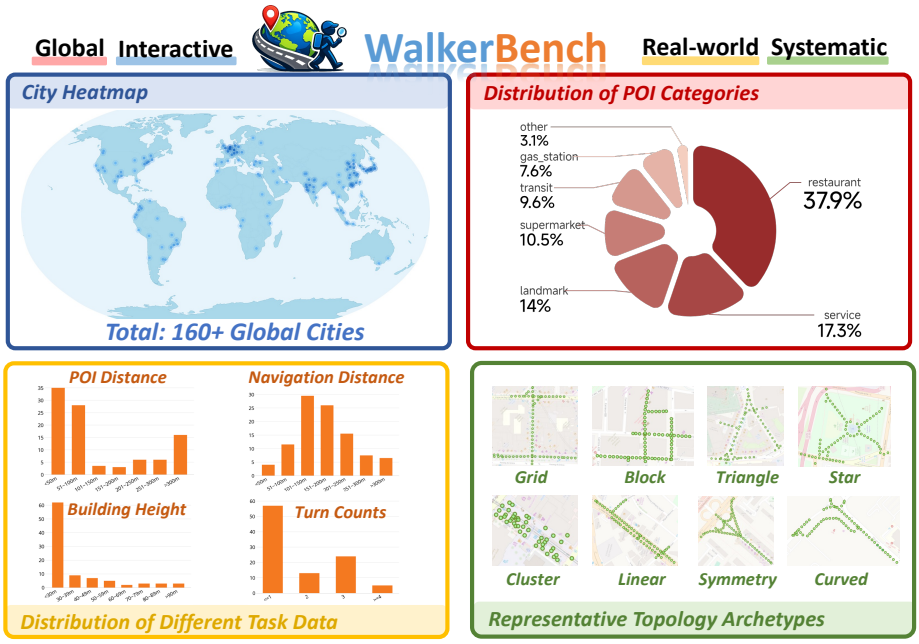

Fig. 2: WalkerBench overview: geographic distribution across 6 continents and 161 cities, POI category composition, task statistics, and the 8 topology archetypes.

Vision-and-Language Navigation. VLN requires agents to follow natural language instructions through visual environments [2, 7], with extensions to multilingual [14] and continuous settings [13]. Recent advances include zero-shot navigation [35, 36] and long-horizon efficiency [5, 28]. However, existing work remains confined to indoor simulators with manual annotation and limited geographic diversity. Ours extends VLN to a global-scale outdoor setting with automatic annotation, while *Spatial-IDE* addresses linear context accumulation via explicit topological memory.

Embodied Environments & Agentic Frameworks. High-fidelity simulators [20, 22, 29] have advanced indoor navigation but face scalability constraints. Outdoor environments built on Google Street View [4,19] are geographically limited and annotation-intensive. Meanwhile, agentic frameworks [26,32,33] demonstrate that complex tasks require specialized scaffolding beyond generic prompting. Ours overcomes geographic and annotation barriers via multi-source cross-validation, and *Spatial-IDE* contributes a topology-aware memory scaffold tailored for embodied navigation.

3 WalkerBench

Evaluating embodied spatial intelligence requires more than a controlled lab—it requires the world. Existing benchmarks confine agents to synthetic environments or pre-rendered trajectories, where static priors can substitute for genuine

reasoning. WalkerBench breaks this constraint by grounding every task in live Google Street View imagery: the only way to acquire spatial knowledge is to move through the environment and observe it. Spanning 6 continents, 161 cities, and 1,000 carefully curated tasks, it probes two cognitive categories—**active perception** and **spatial navigation**—across eight topologically distinct environment archetypes, composing the most geographically and structurally diverse benchmark for embodied agents to date. In the following, we provide a high-level overview of the benchmark.

3.1 Environment Design

The environment is modelled as an implicit graph $\mathcal{E} = (V, E)$ of Google Street View panoramas that is never exposed to the agent. At each step, the agent receives only a first-person 1024×768 RGB image and a relative bearing; GPS coordinates and depth are withheld. The action space is limited to spatial displacement, viewpoint rotation, and Stop. This minimal interface is a deliberate design choice: by denying the agent any privileged geometric signal, we ensure that spatial understanding must emerge from vision alone, mirroring the perceptual constraints a human pedestrian actually faces.

3.2 What Makes WalkerBench Distinctive

Real-world scale and diversity. The richness of WalkerBench stems from its breadth. Tasks are distributed across 7 semantic POI categories—from everyday venues like restaurants (food & dining, 37.9%) to culturally significant landmarks (14.0%)—and sampled from cities on every major continent. This deliberate geographic spread forces agents to generalise across architectural styles, signage conventions, and environmental aesthetics that no synthetic dataset can replicate. Performing well on WalkerBench therefore requires robust cross-domain perception, not pattern matching to a familiar visual vocabulary.

Controlled topology archetypes. Diversity of place alone is not sufficient for systematic evaluation. To isolate *why* an agent succeeds or fails, WalkerBench organises environments into eight graph-theoretic archetypes—Grid, Block, Star, Curved, and four others—characterised by node-degree distribution, clustering coefficient, and spatial spread ratio. Each archetype stresses a distinct facet of spatial cognition: grid layouts demand dead-reckoning over long horizons, star topologies test hub-and-spoke memory, and curved paths challenge bearing estimation. This controlled structure transforms WalkerBench from a leaderboard into a diagnostic instrument.

Rigorous data integrity. Scale and diversity are only valuable if the underlying tasks are trustworthy. An automated multi-stage pipeline (described below) enforces visual solvability, metric accuracy, and contamination resistance at every step, distilling 2,408 raw candidates into 1,000 tasks where each challenge has a verifiable, unambiguous answer.

3.3 Automated Data Engine

Building a benchmark that is simultaneously global in scale, topologically diverse, and free of data contamination demands a purpose-built data pipeline. Our automated engine distils 2,408 raw candidates into 1,000 high-quality, solvable tasks through a cascade of geometric, perceptual, and semantic checks.

Scene coverage begins with an *isotropic BFS* that expands outward from each target POI in eight balanced directional sectors, yielding spatially circular panorama graphs rather than road-biased corridors. Connectivity gaps in Google's native graph are patched with *virtual links*—bidirectional edges added between any two reachable nodes within 18 m—so that the agent's action space faithfully mirrors physical reachability. Spawn points are then drawn via greedy farthest-point sampling, ensuring each task presents multiple distinct approach directions.

Quality is enforced at two levels. A geometric filter removes tasks with absent or outdated Street View coverage, image artifacts, undersized local graphs, or implausible metric values. A subsequent VLM verification pass (Gemini Pro Vision) replays the optimal-path frame sequence and confirms the target POI is visually unambiguous at the final step—excluding any task where the goal is occluded or otherwise indiscernible.

Finally, contamination is suppressed by stripping street names from navigation instructions, sampling non-landmark structures for metric tasks, and drawing heavily from non-English-speaking cities (e.g., Dhaka, Luanda, Karachi) where memorised geographic priors are weakest. The result is a benchmark that demands live visual reasoning rather than recalled world knowledge.

3.4 Task Taxonomy

WalkerBench tasks fall into two complementary categories that together span the full arc of spatial cognition—from reading the environment at a glance to traversing it over many steps.

Category I — Active Perception. Estimating metric properties such as distance, building height, or relative bearing from a single viewpoint is geometrically underdetermined. To resolve this ambiguity, agents must move—accumulating multi-view parallax and integrating observations across time in the way a skilled navigator naturally would. Ground truth for all perception tasks is derived from authoritative geospatial APIs (Google Places, Solar, Elevation), ensuring that evaluation is grounded in objective reality rather than human annotation.

Category II — Spatial Navigation. Navigation tasks ask agents to reach a target POI from a distant spawn point, evaluated by success rate (SR) and success weighted by path length (SPL). Two sub-tasks probe complementary reasoning pathways: **Geometric Navigation** provides turn-by-turn instructions with street names stripped, demanding dead-reckoning from structural cues

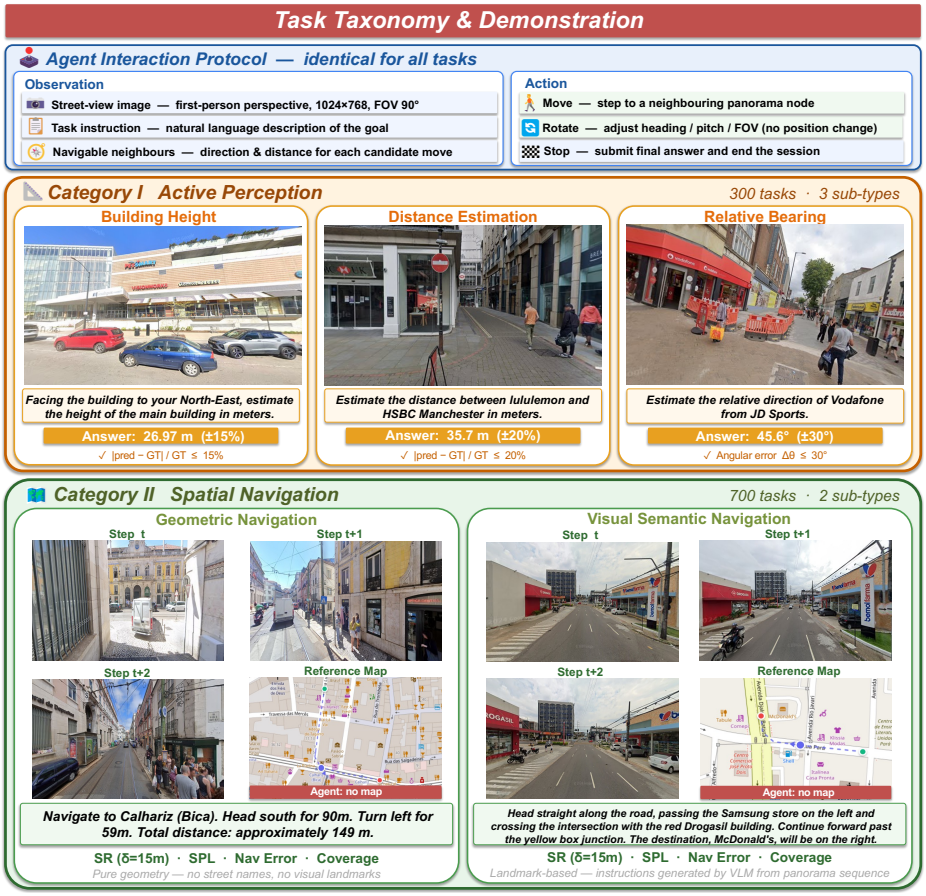

Fig. 3: Task taxonomy and representative instances.

alone; **Visual Semantic Navigation** replaces instructions with landmark-only descriptions generated by a VLM, demanding cross-modal grounding between language and visual scene. Evaluated across all eight topology archetypes, this dual design provides a fine-grained window into where spatial reasoning breaks down—and why.

4 The Spatial-IDE Framework

Spatial-IDE is built around one principle: *Break the Linear Context*. All persistent state is externalised into a structured memory outside the model's context window, and the monolithic per-step call is decomposed into two stateless, single-turn invocations. Every VLM call therefore operates on a *fixed, bounded* input regardless of episode length—no history, no stale observations, no compounding attention drift. Here we give a high-level overview of the framework.

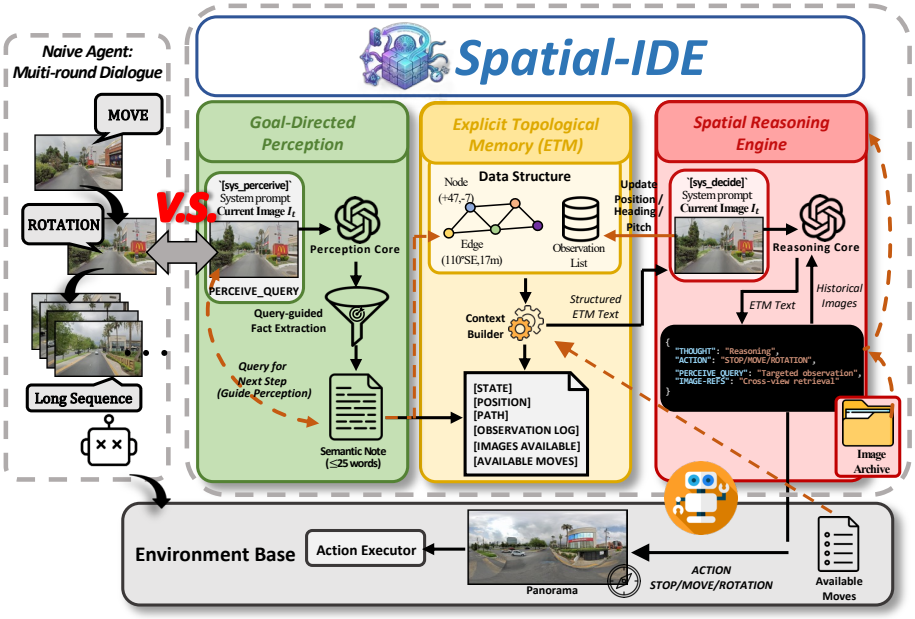

Fig. 4: System architecture of Spatial-IDE. At each step, two stateless VLM calls are mediated by the Explicit Topological Memory (ETM): the *Perceive* call extracts goal-directed visual facts from the current panorama, while the *Decide* call receives a fixed-size serialisation of the ETM and produces the next action with a query steering the next perception pass. The ETM is the sole persistent state; no conversational history is passed to either call.

4.1 Goal-Directed Visual Perception

Rather than open-ended scene narration, the *Goal-Directed Visual Perception* module performs *targeted interrogation*: instead of "what do you see?", it poses a precise, decision-relevant question formulated by the reasoning engine in the previous step—e.g., "Is the pharmacy sign visible, and at what bearing?" The call returns only a compact *Semantic Note*: a short, factual proposition stripped of narrative.

This closes the loop between reasoning and perception—the agent first decides *what to look for*, then looks. Because the perceptual call carries only the current image and a single query, it remains stateless and incurs identical cost at every step.

4.2 Explicit Topological Memory

The Explicit Topological Memory (ETM) is the agent's sole persistent data structure: a dynamic graph $\mathcal{G}_t$ whose nodes record visited viewpoints with absolute coordinates and camera orientation, whose edges encode physical displacement and bearing, and whose per-node annotations accumulate Semantic Notes.

Table 2: Structural comparison: linear-context baseline vs. Spatial-IDE.

Dimension	Linear-Context Baseline	Spatial-IDE
Context growth	Unbounded, step-linear	Fixed per call
Spatial memory	Implicit (attention weights)	Explicit graph $\mathcal{G}_t$
Perception style	Generic scene description	Query-directed extraction
Multi-view reasoning	Implicit / accidental	On-demand & geometric

Crucially, $\mathcal{G}_t$ is a *program-level* structure, not a language model context. Positions are stored in integer-metre coordinates and never drift, fade, or corrupt—the topological model is as reliable at step fifty as at step five. At each decision step, the ETM is projected into a compact, fixed-length *working memory snapshot* that encodes current position, traversal path, and past observations within a constant token budget, ensuring the reasoning engine always receives uniformly high-quality signal.

4.3 Spatial Reasoning Engine

The *Decide* call receives only the current panorama and the serialised ETM snapshot—no prior images, no prior turns. Reasoning quality is governed entirely by the state representation, not by extracting signal from noisy history. Beyond selecting the next action, the engine simultaneously formulates the `PERCEIVE_QUERY` for the *next* viewpoint, making the perception–decision loop genuinely closed.

The engine also supports *on-demand retrieval* of specific historical frames for multi-view geometric reasoning. Because the ETM records coordinates and camera angles for every node, the engine can identify the most geometrically complementary past viewpoint and request exactly that image—surgical and principled, incurring cost only when geometric evidence warrants it.

Table 2 summarises the key contrasts between Spatial-IDE and the linear-context baseline.

5 Experiments

5.1 Implementation Details

Benchmark configuration. WalkerBench covers 160+ cities worldwide and comprises two progressive task tiers: *Tier-I* coarse-grained spatial tasks (navigation **Nav** and visual localisation **Vis**) and *Tier-II* fine-grained metric tasks (height estimation **Height**, distance estimation **Dis**, and angular bearing estimation **Angle**). Each sub-task is scored by accuracy (%), and the **Total** score is the macro-average across all five sub-tasks. Every scene is constructed from real-world street-view panoramas and geo-referenced map tiles; no synthetic data is used.

Table 3: Detailed sub-task and overall performance of vision-language agents with and without **Spatial-IDE**, together with the total-score improvement relative to the original setting. Sub-task scores and Total are reported in accuracy (%). Human (Reference) serves as the upper-bound baseline.

Agent	Setting	Sub-Task Accuracy (%)					Total	Δ Impr.
		Nav	Vis	Height	Dis	Angle		
claude-opus-4-6	Original	52.86	18.57	14.0	9.0	18.0	24.49	—
	w/ Spatial-IDE	**68.29**	**38.57**	**35.0**	**38.0**	**36.0**	**43.17**	**+76.28%**
kimi-k2.5	Original	54.86	9.71	17.0	12.0	15.0	21.71	—
	w/ Spatial-IDE	**69.14**	**28.57**	**37.0**	**31.0**	**33.0**	**39.74**	**+83.05%**
gemini-3-pro-preview	Original	50.86	10.29	12.0	16.0	11.0	20.03	—
	w/ Spatial-IDE	**65.43**	**29.14**	**33.0**	**35.0**	**28.0**	**38.11**	**+90.26%**
gemini-3-flash-preview	Original	44.57	7.43	17.0	17.0	5.0	18.20	—
	w/ Spatial-IDE	**58.86**	**26.29**	**37.0**	**36.0**	**22.0**	**36.03**	**+97.97%**
qwen3-vl-235b-instruct	Original	16.00	10.86	46.0	1.0	13.0	17.37	—
	w/ Spatial-IDE	**40.57**	**30.00**	**58.0**	**20.0**	**31.0**	**35.91**	**+106.74%**
gpt-5-chat-latest	Original	20.00	10.86	16.0	19.0	20.0	17.17	—
	w/ Spatial-IDE	**45.14**	**29.71**	**37.0**	**38.0**	**39.0**	**37.77**	**+119.98%**
claude-sonnet-4-6	Original	33.43	13.71	3.0	18.0	15.0	16.63	—
	w/ Spatial-IDE	**50.00**	**32.57**	**24.0**	**37.0**	**33.0**	**35.31**	**+112.33%**
qwen3-vl-8b-instruct	Original	10.29	10.00	39.0	6.0	8.0	14.66	—
	w/ Spatial-IDE	**32.29**	**28.57**	**52.0**	**23.0**	**25.0**	**32.17**	**+119.44%**
glm-4.6v	Original	29.14	11.43	1.0	13.0	12.0	13.31	—
	w/ Spatial-IDE	**46.57**	**30.29**	**21.0**	**31.0**	**30.0**	**31.77**	**+138.69%**
Average (Original)		*34.67*	*11.43*	*18.33*	*12.33*	*13.00*	*18.17*	—
Average (w/ Spatial-IDE)		*52.92*	*30.41*	*37.11*	*32.11*	*30.78*	*36.66*	**+104.97%**
Human (Reference)	Baseline	78.29	68.86	73.0	66.0	61.0	69.43	—

Agents under evaluation. We evaluate nine vision-language agents spanning a broad capability spectrum. *Proprietary models*: GPT-5-chat-latest [1], Gemini-3-pro-preview [26], Gemini-3-flash-preview [26], Claude-Sonnet-4.6 , and Claude-Opus-4.6. *Open-source models*: GLM-4.6V [11], Kimi-K2.5 [27], Qwen3-VL-8B-Instruct [3], and Qwen3-VL-235B-A22B-Instruct [3]. Each agent is tested in two settings: *Original* (standard multi-turn dialogue with raw image observations) and *w/ Spatial-IDE* (our proposed framework, described in Section 4). Performance from five human annotators is included as a human reference.

5.2 Main Results

Overview. Table 3 reports sub-task and total accuracy for every agent in both settings. Three findings stand out immediately. First, *Spatial-IDE yields consistent, large improvements for every agent without exception*: the per-model gain ranges from +76.28% (Claude-Opus-4.6) to +138.69% (GLM-4.6V), with an average of +**104.97**%. Second, *the absolute ranking of agents is partially reshuffled* by Spatial-IDE, indicating that the original ranking conflates raw visual capability with the ability to manage spatial context—two properties our framework deliberately decouples. Third, *a sizable gap to the human reference (69.43%) remains*, demonstrating that WalkerBench is not yet saturated and provides a meaningful long-horizon challenge for future work.

Sub-task analysis. Examining the five sub-tasks reveals distinct patterns.

Navigation (Nav) is the strongest sub-task for nearly all models in the original setting (average 34.67%), because coarse way-finding can often be resolved from a single panoramic observation. Spatial-IDE raises the average to 52.92%—a +52.7% relative gain—by ensuring that previously visited nodes are not "forgotten" when the agent accumulates more than a handful of moves.

Visual localisation (Vis) is the weakest sub-task under the original setting (average 11.43%), confirming our hypothesis that passive single-image representations are fundamentally insufficient for grounding in a non-Euclidean street topology. Spatial-IDE more than doubles the average score to 30.41% (+166.1% relative), the largest proportional gain of any sub-task. This suggests that structuring observations into an explicit topological map is especially critical for cross-view correspondence.

Metric sub-tasks (Height, Dis, Angle) all start from low baselines ($\leq$18% on average) and reach 37.11%, 32.11%, and 30.78% respectively under Spatial-IDE. The gains here are attributable primarily to Goal-Directed Perception: by isolating task-relevant visual cues (e.g., façade lines for height, shadow geometry for angle) before passing them to the reasoning chain, the VLM is relieved of having to simultaneously manage spatial memory and low-level measurement, two cognitively competing demands.

Proprietary vs. open-source models. Among proprietary models, Claude-Opus-4.6 achieves the highest absolute total score in both the original (24.49%)

and the Spatial-IDE (43.17%) settings. Notably, GPT-5-chat-latest starts with the *lowest* total among proprietary models (17.17%) yet achieves the second-largest relative improvement (+119.98%), reaching 37.77% with Spatial-IDE. This suggests that GPT-5's spatial deficiencies under the original setting stem largely from context-management failure rather than insufficient perceptual capacity—exactly the failure mode Spatial-IDE is designed to address.

Among open-source models, Qwen3-VL-235B-Instruct and Kimi-K2.5 are competitive with mid-tier proprietary models under Spatial-IDE (35.91% and 39.74%, respectively), closing much of the capacity gap that is visible in the original setting. The 8B-parameter Qwen3-VL-8B-Instruct reaches 32.17% with Spatial-IDE, a result that was entirely out of reach without the framework (14.66%), demonstrating that Spatial-IDE is particularly beneficial for smaller models whose context windows are more easily overwhelmed.

Effect of model scale. Comparing Qwen3-VL-8B-Instruct and Qwen3-VL-235B-Instruct, the larger model consistently leads in absolute accuracy (35.91% vs. 32.17% with Spatial-IDE), as expected. However, both models improve by a similar *relative* margin (+106.74% vs. +119.44%), indicating that the representational mis-alignment between linear context and three-dimensional topology—which we term *spatial forgetting*—is a structural property of current transformer architectures rather than a scaling artefact.

Error ceiling and human gap. Even the best agent with Spatial-IDE (Claude-Opus-4.6 at 43.17%) lags the human reference by 26.26 percentage points. The gap is largest in **Vis** (68.86% human vs. 38.57% best agent), highlighting that fine-grained cross-view grounding remains an open research challenge. The relatively smaller gap in **Height** (73.0% human vs. 58.0% best agent with Spatial-IDE) aligns with the observation that metric estimation from single images is a well-studied problem where VLMs have absorbed relevant priors during pre-training.

5.3 Ablation Study

To rigorously attribute the performance gains of Spatial-IDE, we conduct a systematic ablation over its two core components: **Explicit Topological Memory (ETM)** and **Goal-Directed Perception (GDP)**. We construct four experimental conditions—Original, ETM-only, GDP-only, and the full Spatial-IDE (ETM+GDP)—and report scores averaged across all nine agents on the complete WalkerBench test set. Results are summarised in Table 4.

ETM addresses the spatial-memory bottleneck. Adding ETM alone (row 2) raises the total from 18.17% to 27.83% (+9.66 points). The gain is most pronounced on **Nav** (+12.47 points), the sub-task most sensitive to long-horizon spatial coherence: without ETM, earlier observations are effectively forgotten as

Table 4: Ablation study on the two core components of Spatial-IDE. All numbers are averaged across all nine agents. Sub-task scores and Total are reported in accuracy (%). Δ is the absolute improvement over the Original baseline.

Setting	Nav	Vis	Height	Dis	Angle	Total	Δ
Original (baseline)	34.67	11.43	18.33	12.33	13.00	18.17	—
+ ETM only	47.14	21.00	25.00	23.00	23.00	27.83	+9.66
+ GDP only	39.00	26.57	33.00	26.57	25.57	30.14	+11.97
+ ETM + GDP (Full)	**52.92**	**30.41**	**37.11**	**32.11**	**30.78**	**36.66**	**+18.49**

trajectory length grows. ETM converts transient observations into a persistent adjacency graph, enabling $O(1)$ neighbourhood retrieval regardless of trajectory length. The modest gains on metric sub-tasks (Height, Dis, Angle) confirm that ETM helps with structural memory, not fine-grained perceptual measurement.

GDP addresses the attentional bottleneck. Adding GDP alone (row 3) raises the total to 30.14% (+11.97 points), with gains distributed differently: **Vis** jumps by +15.14 points and each metric sub-task gains over +12 points, while Nav improves only modestly (+4.33 points). This pattern reflects GDP's design: a lightweight task-conditioned detection head isolates cues relevant to the current sub-task, eliminating the attentional overhead of processing irrelevant scene content. The weaker Nav gain confirms that navigation is primarily a memory problem—precisely the failure mode ETM is designed to fix.

ETM and GDP are synergistic. The full Spatial-IDE framework (row 4) achieves 36.66% (+18.49 points over baseline), exceeding a naive additive prediction and indicating moderate positive interaction between components. The effect is most visible on **Vis**, where joint gains surpass simple additivity: reliable cross-view grounding requires both structured memory of prior views *and* a clean perceptual signal for each new view. Together, ETM and GDP decompose the spatial-reasoning challenge into two well-scoped sub-problems—*what to remember* and *what to look at*—allowing VLMs to solve them without being overwhelmed by their combination.

5.4 Zero-Shot Transfer to Physical Urban Navigation

The ultimate test of any spatial intelligence framework is its ability to guide an embodied agent through the complexity of the real world. To this end, we deploy Spatial-IDE on a **Unitree G1** humanoid robot navigating open urban environments, demonstrating that the spatial reasoning capabilities acquired on WalkerBench transfer directly to physical city-scale navigation.

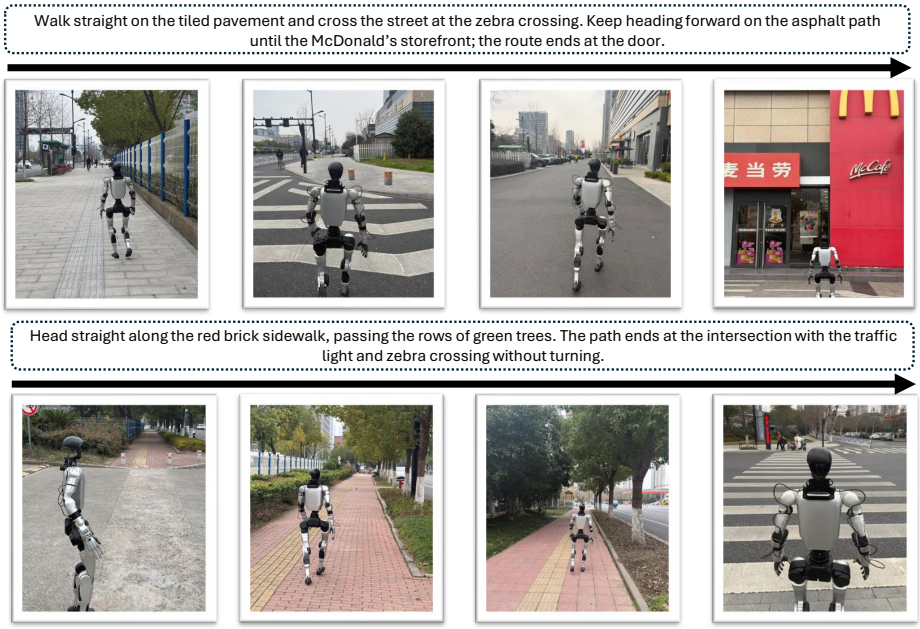

Fig. 5: Real World Deployment of Spatial-IDE on a Unitree G1 Robot.

System Overview. As illustrated in Figure 5, the robot receives natural-language route instructions and must execute long-distance, multi-waypoint navigation across city blocks entirely autonomously. Low-level locomotion is handled by the **Unitree official motion controller**, which abstracts bipedal walking and stability from the high-level planner. At each decision step, a dedicated **VLM-based traversability detector** identifies locally navigable candidates in the robot's field of view, supplying a set of actionable waypoints to Spatial-IDE's reasoning module. The high-level planner then selects among these candidates according to the Explicit Topological Memory (ETM) and the goal-directed perception mechanism described in Section 4.

Long-Distance Urban Navigation. The decoupled architecture of Spatial-IDE imposes no assumption on observation modality beyond a standard RGB image stream, making it immediately portable to the robot's onboard camera. The agent successfully navigates multi-block routes spanning semantically diverse urban scenes—tiled sidewalks, zebra crossings, asphalt roads, and landmark storefronts—guided solely by free-form language instructions of the type shown in Figure 5. These results demonstrate that the structured spatial reasoning instilled by Spatial-IDE generalises zero-shot from web-sourced street-view data to genuine embodied deployment.

6 Conclusion

In this work, we identify and address the critical "representation misalignment" between the one-dimensional linear context of Vision-Language Models (VLMs) and the topological nature of 3D environments. To overcome the resulting spatial forgetting, we introduce *WalkerBench*, a global-scale interactive benchmark for spatial intelligence, and propose *Spatial-IDE*. By leveraging State Externalization to build an Explicit Topological Memory (ETM) and employing Cognitive Decoupling, our framework enables VLMs to focus on high-level decision-making. Spatial-IDE not only achieves substantial improvements on WalkerBench but also successfully generalizes zero-shot to real-world urban navigation on a humanoid robot, providing a robust foundation for active physical interaction in embodied agents.

7 Acknowledgement

This work was supported by the National Key R&D Program of China (2025ZD0123100), the NSFC (62436007), the Key R&D Program in Zhejiang Province (2026SDXT005, 2025C01030), the Zhejiang NSF (LQK26F020001, LRG25F020001), the Ministry of Culture and Tourism (No.2023DMKLB002).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3674–3683 (2018)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Chen, H., Suhr, A., Misra, D., Snavely, N., Artzi, Y.: Touchdown: Natural language navigation and spatial reasoning in visual street environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12538–12547 (2019)
5. Cheng, A.C., Ji, Y., Yang, Z., Gongye, Z., Zou, X., Kautz, J., Bıyık, E., Yin, H., Liu, S., Wang, X.: Navila: Legged robot vision-language-action model for navigation. arXiv preprint arXiv:2412.04453 (2024)
6. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
7. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. Advances in neural information processing systems **31** (2018)

8. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
9. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
10. Gupta, A., Savarese, S., Ganguli, S., Fei-Fei, L.: Embodied intelligence via learning and evolution. Nature communications **12**(1), 5721 (2021)
11. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
12. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
13. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)
14. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. arXiv preprint arXiv:2010.07954 (2020)
15. Li, Y., Zhang, Y., Lin, T., Liu, X., Cai, W., Liu, Z., Zhao, B.: Sti-bench: Are mllms ready for precise spatial-temporal world understanding? arXiv preprint arXiv:2503.23765 (2025)
16. Lin, J., Xu, R., Zhu, S., Yang, S., Cao, P., Ran, Y., Hu, M., Zhu, C., Xie, Y., Long, Y., et al.: Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. arXiv preprint arXiv:2512.10863 (2025)
17. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
18. Mehta, H., Artzi, Y., Baldridge, J., Ie, E., Mirowski, P.: Retouchdown: Adding touchdown to streetlearn as a shareable resource for language grounding tasks in street view. arXiv preprint arXiv:2001.03671 (2020)
19. Mirowski, P., Banki-Horvath, A., Anderson, K., Teplyashin, D., Hermann, K.M., Malinowski, M., Grimes, M.K., Simonyan, K., Kavukcuoglu, K., Zisserman, A., et al.: The streetlearn environment and dataset. arXiv preprint arXiv:1903.01292 (2019)
20. NVIDIA: Isaac sim-robotics simulation and synthetic data generation (2023), `https://developer.nvidia.com/isaac/sim`
21. Roy, N., Posner, I., Barfoot, T., Beaudoin, P., Bengio, Y., Bohg, J., Brock, O., Depatie, I., Fox, D., Koditschek, D., et al.: From machine learning to robotics: Challenges and opportunities for embodied intelligence. arXiv preprint arXiv:2110.15245 (2021)
22. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9339–9347 (2019)
23. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15768–15780 (2025)

24. Sumers, T., Yao, S., Narasimhan, K.R., Griffiths, T.L.: Cognitive architectures for language agents. Transactions on Machine Learning Research (2023)
25. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
26. Team, G.R., Abdolmaleki, A., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Balakrishna, A., Batchelor, N., Bewley, A., Bingham, J., et al.: Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. arXiv preprint arXiv:2510.03342 (2025)
27. Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., et al.: Kimi k2: Open agentic intelligence. arXiv preprint arXiv:2507.20534 (2025)
28. Wei, M., Wan, C., Yu, X., Wang, T., Yang, Y., Mao, X., Zhu, C., Cai, W., Wang, H., Chen, Y., et al.: Streamvln: Streaming vision-and-language navigation via slowfast context modeling. arXiv preprint arXiv:2507.05240 (2025)
29. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9068–9079 (2018)
30. Xu, P., Wang, S., Zhu, Y., Li, J., Zhang, Y.: Spatialbench: Benchmarking multimodal large language models for spatial cognition. arXiv preprint arXiv:2511.21471 (2025)
31. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
32. Yang, J., Jimenez, C.E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., Press, O.: Swe-agent: Agent-computer interfaces enable automated software engineering. Advances in Neural Information Processing Systems **37**, 50528–50652 (2024)
33. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
34. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. arXiv preprint arXiv:2504.15280 (2025)
35. Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: Navid: Video-based vlm plans the next step for vision-and-language navigation. arXiv preprint arXiv:2402.15852 (2024)
36. Zhou, G., Hong, Y., Wu, Q.: Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7641–7649 (2024)
37. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, X.E., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8600–8612 (2025)